Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 97 (2006) 231–245

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Adaptation under probabilistic error for estimating linear functionals☆

T. Tony Cai*, Mark G. Low

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

Received 4 November 2003
Available online 2 April 2005

Abstract

The problem of estimating linear functionals based on Gaussian observations is considered. Probabilistic error is used as a measure of accuracy and attention is focused on the construction of adaptive estimators which are simultaneously near optimal under probabilistic error over a collection of convex parameter spaces. In contrast to mean squared error it is shown that fully rate optimal adaptive estimators can be constructed for probabilistic error. A general construction of such estimators is provided and examples are given to illustrate the general theory.

AMS 1991 subject classification: primary 62G99; secondary 62F12; 62F35; 62M99

Keywords: Adaptive estimation; Confidence intervals; Gaussian models; Modulus of continuity; Probabilistic error

1. Introduction

One of the central goals in nonparametric function estimation is the development of procedures which adapt to the smoothness of the underlying function. One way to formalize this goal is to focus on the construction of procedures which are simultaneously near minimax over a collection of parameter spaces. Estimators which attain such a goal are termed adaptive.

☆ Research supported in part by NSF Grant DMS-0306576.

* Corresponding author.
E-mail address: tcai@wharton.upenn.edu (T. Tony Cai).

0047-259X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2005.03.001

Adaptation theory depends critically on how risk is measured. For estimating linear functionals the problem of adaptation has mostly been studied under mean squared error. For estimating a function at a point, Lepski [19] was the first to show in a number of cases that under mean squared error the exact minimax rate of convergence cannot be attained simultaneously over even a pair of Lipschitz classes. A logarithmic penalty must be paid for adaptation. Further developments can be found for example in [20,22,15]. Efromovich and Low [12] extended pointwise estimation to arbitrary linear functionals over symmetric convex parameter spaces.

A more general theory for adaptation under mean squared error has been developed in [3]. This theory covers general convex parameter spaces using geometric quantities, namely the ordered and between class moduli of continuity. For a linear functional T and parameter spaces $\mathcal{F}$ and $\mathcal{G}$ the ordered modulus of continuity $\omega(\varepsilon, \mathcal{F}, \mathcal{G})$ is defined by

$$\omega(\varepsilon, \mathcal{F}, \mathcal{G}) = \sup\{Tg - Tf : \|g - f\|_2 \leqslant \varepsilon;\ f \in \mathcal{F},\ g \in \mathcal{G}\}. \tag{1}$$

When $\mathcal{G} = \mathcal{F}$, $\omega(\varepsilon, \mathcal{F}, \mathcal{F})$ is the modulus of continuity over $\mathcal{F}$, introduced by Donoho and Liu [9] for minimax theory, and will be denoted by $\omega(\varepsilon, \mathcal{F})$. The between class modulus of continuity is then defined as $\omega_+(\varepsilon, \mathcal{F}, \mathcal{G}) = \max\{\omega(\varepsilon, \mathcal{F}, \mathcal{G}), \omega(\varepsilon, \mathcal{G}, \mathcal{F})\}$ or equivalently

$$\omega_+(\varepsilon, \mathcal{F}, \mathcal{G}) = \sup\{|Tg - Tf| : \|g - f\|_2 \leqslant \varepsilon;\ f \in \mathcal{F},\ g \in \mathcal{G}\}. \tag{2}$$

These moduli were shown to be instrumental in characterizing the possible degree of adaptability over two convex classes $\mathcal{F}$ and $\mathcal{G}$ in the same way that the modulus of continuity $\omega(\varepsilon, \mathcal{F})$ was used by Donoho and Liu [9] and Donoho [8] to capture the minimax risk over a single convex parameter space $\mathcal{F}$. In particular the general theory given in [3] shows that sometimes the cost of adaptation can be much more than a logarithmic factor.

In addition to mean squared error there are other natural performance measures. In this paper we shall focus on probabilistic error. For a given estimator $\hat{T}$, linear functional Tf and precision level c the probabilistic error over $\mathcal{F}$ is given by

$$\sup_{f \in \mathcal{F}} P(|\hat{T} - Tf| \geqslant c). \tag{3}$$

It is often natural, as in the construction of confidence intervals to first specify a tolerance level α for the probabilistic error and seek the smallest value of c in (3) attaining such a level. For a given α the minimax benchmark is given by

$$r(\mathcal{F}, \alpha) = \inf\left\{d : \inf_{\hat{T}} \sup_{f \in \mathcal{F}} P(|\hat{T} - Tf| \geqslant d) \leqslant \alpha\right\}. \tag{4}$$

Asymptotic versions of (4) are one way to define an optimal rate of convergence. See for example [23,14,8]. Note also that $r(\mathcal{F}, \alpha)$ is the half length of the shortest fixed length confidence interval with coverage probability over $\mathcal{F}$ of at least $1-\alpha$. Following an approach of Donoho [8] lower bounds for $r(\mathcal{F}, \alpha)$ are given in Section 2 for arbitrary parameter spaces. An alternative approach not considered in this paper is to fix a precision level c in (3) and to construct an optimal procedure $\hat{T}$ minimizing the probabilistic error over $\mathcal{F}$ given in (3). The asymptotic behavior of this Bahadur risk has for example been studied in [17].

In this paper, we are primarily interested in adaptive estimation under probabilistic error over a collection of parameter spaces. There is a dramatic difference between mean squared error adaptation and probabilistic error adaptation. In contrast to mean squared error where the minimax rate of convergence cannot in general be attained simultaneously over even two parameter spaces it is shown in Section 3 that under probabilistic error fully rate optimal adaptation is always possible over any finite collection of convex parameter spaces. A general construction of a rate optimal adaptive estimator is given based on the ordered modulus of continuity. The adaptive estimator is constructed using tests between parameter spaces which are based on estimators which trade bias and variance in a precise way. This general approach was used in [19] but the specific tests are designed for probabilistic error.

For a collection of k convex parameter spaces upper bounds for the performance of this estimator are given in terms of k and the moduli of continuity. Illustrative examples are given in Section 4. One example considers Lipschitz classes which contrast the difference between mean squared error and probabilistic error adaptation when k is fixed. An example is also given where the number of convex parameter spaces is not fixed and which shows that the upper bound is rate optimal as a function of k.

2. Minimax theory

Throughout this paper we consider estimation of a linear functional Tf based on Gaussian observations

$$dY(t) = f(t)\,dt + \frac{1}{\sqrt{n}}\,dW(t), \tag{5}$$

where W is standard Brownian motion or

$$Y(i) = f(i) + \frac{1}{\sqrt{n}}\,z_i, \tag{6}$$

where z_i are i.i.d. standard normal random variables.

In this context minimax theory is well developed for squared error loss. In particular one of the central results of Donoho and Liu [9] is that for any linear functional and a given convex parameter space the associated modulus of continuity determines the minimax rate of convergence under a large number of loss functions. For example under squared error loss and Gaussian observations $\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F})$ is the minimax rate for estimating the linear functional Tf over a convex parameter space $\mathcal{F}$. Extensions of these results to nonconvex parameter spaces has been given in [6].

Results have also been developed for the shortest fixed length confidence intervals with a given level of coverage. See [8]. These results have immediate implications for probabilistic error. In particular it follows from the confidence interval results in [8] that for any $\alpha > 0$ and a given convex parameter space $\mathcal{F}$ there exists a linear estimator $\hat{T}$ such that

$$\sup_{f\in\mathcal{F}} P\left(|\hat{T} - Tf| > \omega\left(\frac{2z_{\alpha/2}}{\sqrt{n}}, \mathcal{F}\right)\right) \leqslant \alpha \tag{7}$$

and also that for any procedure $\hat{T}$

$$\sup_{f\in\mathcal{F}} P\left(|\hat{T}-Tf| > \omega\left(\frac{2z_\alpha}{\sqrt{n}},\mathcal{F}\right)\right) > \alpha. \tag{8}$$

The focus of this paper is on adaptation theory under probabilistic error. For the development of such a theory it is necessary to derive lower bounds for probabilistic error over arbitrary parameter spaces. The following theorem which essentially follows from ideas in [8] provides such a lower bound.

Theorem 1. *Let $\mathcal{G}$ be a parameter space and let $0 < \alpha < \frac{1}{2}$. For estimating a linear functional Tf based on Gaussian models* (5) *or* (6)

$$\sup_{f\in\mathcal{G}} P\left(|\hat{T}-Tf| \geqslant \frac{1}{2}\omega\left(\frac{2z_\alpha}{\sqrt{n}},\mathcal{G}\right)\right) \geqslant \alpha \tag{9}$$

or equivalently

$$r(\mathcal{G},\alpha) \geqslant \frac{1}{2}\omega\left(\frac{2z_\alpha}{\sqrt{n}},\mathcal{G}\right). \tag{10}$$

Theorem 1 provides a benchmark for the evaluation of any procedure under probabilistic error. For a given estimator $\hat{T}$ let

$$r(\hat{T},\mathcal{G},\alpha) = \inf\left\{d : \sup_{f\in\mathcal{G}} P(|\hat{T}-Tf| \geqslant d) \leqslant \alpha\right\}. \tag{11}$$

Results in Section 3 on adaptive estimation will show that the lower bound given in (10) is rate optimal. The construction in Section 3 also provides a minimax rate optimal estimator for probabilistic error over a finite union of convex sets. More specifically let $\mathcal{G} = \cup_{i=1}^{k} \mathcal{F}_i$ where $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k$ is a collection of convex parameter spaces with pairwise nonempty intersections. Then the estimator $\hat{T}$ given in (16) satisfies

$$\sup_{f\in\mathcal{G}} P\left(|\hat{T}-Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}},\mathcal{G}\right)\right) \leqslant \alpha$$

and hence

$$r(\mathcal{G},\alpha) \leqslant r(\hat{T},\mathcal{G},\alpha) \leqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}},\mathcal{G}\right). \tag{12}$$

This result is a direct consequence of Theorem 2. A comparison of (10) and (12) shows that when the parameter space $\mathcal{G}$ is a finite union of convex sets the minimax rate of convergence under probabilistic error is determined by the modulus of continuity.

3. Adaptation

In this section we turn to the development of an adaptation theory for probabilistic error. Perhaps the most notable feature of this theory is that fully rate adaptive procedures always

exist over a finite collection of convex parameter spaces. This is in sharp contrast to the theory for mean squared error where penalties often must be paid for adaptation. We first briefly recall some known results on adaptation under mean squared error in Section 3.1 and then turn to probabilistic error in Section 3.2.

3.1. Adaptation under mean squared error

The construction of adaptive estimators is a central problem in nonparametric function estimation. For estimating linear functionals the focus so far has been adaptation under mean squared error. Lepski [19] was the first to show that for estimating a function at a point over a collection of Lipschitz classes a logarithmic penalty must be paid for adaptation. Efromovich and Low [12] showed that this phenomenon is in general true for estimating linear functionals over a collection of nested symmetric convex parameter spaces. On the other hand Cai and Low [3] show that in some settings the cost of adaptation can be much more than a logarithmic factor. Under order restrictions such as estimating monotone functions at a point Kang and Low [16] show that fully rate adaptive estimators sometimes exist. Efromovich [11] shows that these results are not just an asymptotic phenomena but are also noticeable in small data sets.

However for integrated squared error it is well known that rate optimal adaptive estimators can often be constructed. In fact it was first shown in [13] that sharp adaptive estimators can be constructed over collections of Sobolev spaces. These estimators are simultaneously asymptotically minimax over each Sobolev ball in the collection. Cai and Low [7] considers the problem of adaptation over shrinking neighborhoods which includes as special cases estimation at a point and estimation over the whole interval. It is shown that the cost of adaptation depends critically on the size of the neighborhood over which the risk is measured.

3.2. Adaptation under probabilistic error

We shall now turn to the problem of adaptation under probabilistic error. The adaptation problem is formulated as follows. Let $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k$ be a collection of convex parameter spaces with pairwise nonempty intersections. Set $\mathcal{G}_j = \cup_{i=1}^{j} \mathcal{F}_i$. In this section we consider adaptation over $\mathcal{G}_j$. Note that $\{\mathcal{G}_j,\ j = 1, \ldots, k\}$ is a collection of nested but not necessarily convex parameter spaces. The benchmark for probabilistic error over each of these spaces is provided by Theorem 1. The goal is to construct a single estimator $\hat{T}$ which simultaneously attains, within a fixed constant factor, the lower bound in (10) for all $\mathcal{G}_j$.

As in the mean squared error adaptation theory in [3] and the confidence interval theory given in [4] the construction of the adaptive procedure relies on the ordered modulus of continuity as given in (1). The details of the construction of the adaptive estimator are however quite different. An important preliminary step is to find linear estimators with precise bias-variance tradeoffs over pairs of parameter spaces. It should be mentioned that the bias tradeoff is given in terms of upper bounds over one parameter space and lower bounds over the other space. These linear estimators can be described as follows. For $1 \leqslant i, j \leqslant k$ set $\omega_{i,j} = \omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j\right)$. The bias-variance tradeoff is then accomplished

by estimators given in [6]. These linear estimators $\hat{T}_{i,j}$ have variance bounded by $\frac{1}{z^2_{\alpha/2k}}\omega^2_{i,j}$ and bias which satisfies

$$\inf_{f\in\mathcal{F}_j}(E(\hat{T}_{i,j})-Tf)\geqslant -\frac{1}{2}\omega_{i,j} \tag{13}$$

and

$$\sup_{f\in\mathcal{F}_i}(E(\hat{T}_{i,j})-Tf)\leqslant \frac{1}{2}\omega_{i,j}. \tag{14}$$

This collection of linear estimators will be used for the construction of a nonlinear adaptive procedure. For convenience write $\hat{T}_i$ for $\hat{T}_{i,i}$. It should be noted that the estimator $\hat{T}_i$ has variance and squared bias bounded above by

$$Var(\hat{T}_i)\leqslant \frac{1}{z^2_{\alpha/2k}}\omega^2\left(\frac{z_{\alpha/2k}}{\sqrt{n}},\mathcal{F}_i\right), \quad \text{and} \quad \sup_{f\in\mathcal{F}_i} Bias^2(\hat{T}_i)\leqslant \frac{1}{4}\omega^2\left(\frac{z_{\alpha/2k}}{\sqrt{n}},\mathcal{F}_i\right).$$

Consequently the estimator $\hat{T}_i$ is minimax rate optimal over $\mathcal{F}_i$ for squared error loss. Similarly for probabilistic error straightforward calculations also yield

$$\sup_{f\in\mathcal{F}_i} P\left(|\hat{T}_i-Tf|>\frac{3}{2}\omega\left(\frac{z_{\alpha/2}}{\sqrt{n}},\mathcal{F}_i\right)\right)\leqslant\alpha. \tag{15}$$

Hence, from the general lower bound (9), $r(\hat{T}_i,\mathcal{F},\alpha)$ is within a small constant factor of the minimax benchmark $r(\mathcal{F},\alpha)$.

The construction of the adaptive procedure is based on a sequence of tests between the different pairs of parameter spaces. For $1\leqslant p\leqslant k-1$ let

$$\Lambda_p=\prod_{p<l\leqslant k} 1\left(\hat{T}_{p,l}-\frac{3}{2}(\omega_{p,l}+\omega_p)\leqslant\hat{T}_p\leqslant\hat{T}_{l,p}+\frac{3}{2}(\omega_{l,p}+\omega_p)\right)$$

and set $\Lambda_0=0$. Λ_p is a test between $\mathcal{F}_p$ and $\cup_{j=p+1}^k\mathcal{F}_j$. The test chooses $\mathcal{F}_p$ with probability of at least $1-\alpha/2k$ when $f\in\mathcal{F}_p$.

The nonlinear adaptive estimator can now be described as follows.

1. Test between $\mathcal{F}_1$ and $\cup_{j=2}^k\mathcal{F}_j$ using Λ_1.
2. If the test Λ_1 chooses $\mathcal{F}_1$ then use $\hat{T}_1$ as the estimate of Tf.
3. Otherwise, delete $\mathcal{F}_1$ and repeat Steps 1 and 2.

More formally the procedure can be written as

$$\hat{T}=\sum_{i=1}^k\hat{T}_i\prod_{j=1}^{i-1}(1-\Lambda_j)\Lambda_i. \tag{16}$$

The following theorem shows that the estimator $\hat{T}$ is fully rate optimally adaptive over the collection $\mathcal{G}_j$.

Theorem 2. *For $1 \leqslant j \leqslant k$ let $\mathcal{G}_j = \cup_{i=1}^{j} \mathcal{F}_i$ where $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k$ is a collection of convex parameter spaces with pairwise nonempty intersections. For all $1 \leqslant j \leqslant k$ the estimator given in* (16) *satisfies*

$$\sup_{f \in \mathcal{G}_j} P_f(|\hat{T} - Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)) \leqslant \alpha$$

and hence

$$r(\hat{T}, \mathcal{G}_j, \alpha) \leqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right).$$

Remark. The estimator given in (16) is specifically designed for adaptation under probabilistic error. It necessarily has poor performance under mean squared error. See [5].

As mentioned earlier for a fixed parameter space confidence intervals can be obtained by inverting the result for probabilistic error. In this setting Theorem 2 immediately yields a result for confidence intervals.

Corollary 1. *Let $\mathcal{G} = \cup_{i=1}^{k} \mathcal{F}_i$ where $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k$ is a collection of convex parameter spaces with pairwise nonempty intersections. Let $\hat{T}$ be given as in* (16). *Then*

$$\hat{T} \pm \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}\right)$$

is a confidence interval for Tf which has coverage probability of at least $1 - \alpha$ over $\mathcal{G}$.

It follows that the length of the optimal fixed length confidence interval with a given level of coverage over a finite union of convex sets is determined up to a constant by the modulus of continuity over the parameter space. Moreover the lower bound given in (9) shows that no random length confidence interval with the same level of coverage can have maximum expected length with a faster rate of convergence.

Sometimes as in the nearly black object example given in Section 4 it is of interest to consider collections of parameter spaces which are unions of a growing number (as n increases) of convex sets. It is then useful to note that $z_{\alpha/2k} \leqslant \sqrt{\frac{2}{z_{\alpha/2}^2}\log k + 1} \cdot z_{\alpha/2}$. Note also that the ordered modulus of continuity $\omega(\varepsilon, \mathcal{H}, \mathcal{J})$ is a concave function of ε whenever both $\mathcal{H}$ and $\mathcal{J}$ are convex parameter spaces with nonempty intersection. See [6]. Hence for $D \geqslant 1$

$$\begin{aligned}\omega(D\varepsilon, \cup_{l=1}^{m} \mathcal{F}_l) &= \max_{1 \leqslant i,j \leqslant m} \omega(D\varepsilon, \mathcal{F}_i, \mathcal{F}_j) \leqslant \max_{1 \leqslant i,j \leqslant m} D\omega(\varepsilon, \mathcal{F}_i, \mathcal{F}_j) \\ &= D\omega(\varepsilon, \cup_{l=1}^{m} \mathcal{F}_l).\end{aligned}$$

It then follows that

$$\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right) \leqslant \sqrt{\frac{2}{z_{\alpha/2}^2}\log k + 1} \cdot \omega\left(\frac{z_{\alpha/2}}{\sqrt{n}}, \mathcal{G}_j\right)$$

and consequently

$$r(\hat{T}, \mathcal{G}_j, \alpha) \leqslant \frac{13}{2}\sqrt{\frac{2}{z_{\alpha/2}^2}\log k + 1} \cdot \omega\left(\frac{z_{\alpha/2}}{\sqrt{n}}, \mathcal{G}_j\right). \tag{17}$$

Comparing the upper bound (17) with the minimax lower bound given in (9) shows that the upper bound contains an extra $\sqrt{\log k}$ factor. The nearly black object example in Section 4 shows that this $\sqrt{\log k}$ factor cannot in general be removed when k grows with n.

The result given in Theorem 2 should be contrasted to results for adaptation under mean squared error and also for adaptive confidence intervals. It is known that under mean squared error fully rate adaptive estimators typically do not exist and a penalty usually must be paid for adaptation. See [3]. Although there is a direct connection between the construction of optimal confidence intervals and minimax results for probabilistic error over a fixed parameter space, there is however a major difference for the corresponding adaptive estimation problems. In the adaptive setting it is not possible to invert the bounds given for probabilistic error since these bounds depend upon the unknown parameter space. There are many examples where it is possible to construct estimators that are guaranteed to be "optimally close" probabilistically to the unknown functional without knowing the accuracy of the estimator. For the construction of confidence intervals the accuracy of the estimator is however needed and consequently it is more difficult to construct adaptive confidence intervals than to construct estimators which are adaptive probabilistically.

4. Examples

In this section we consider two examples to illustrate the results of Section 3. The first example has a fixed number of parameter spaces. It is used to highlight the difference between mean squared error and probabilistic error adaptation. In the second example the number of parameter spaces increases with n. This example is used to show that the factor $z_{\alpha/2k} \sim \sqrt{\log k}$ (for large k) in the upper bound in Theorem 2 cannot in general be dropped.

4.1. Lipschitz classes

Lipschitz classes are one of the most commonly studied parameter spaces in nonparametric function estimation particularly when the object is to recover the function at a given point. For these classes it is well known that under mean squared error a minimum penalty of a logarithmic factor must be paid for adaptation. See [19,2]. We shall see that fully rate optimal adaptation can be achieved over any finite collection of Lipschitz classes under probabilistic error.

For $\beta > 0$ the Lipschitz function class is defined as

$$F(\beta, M) = \left\{ f : \left[-\frac{1}{2}, \frac{1}{2}\right] \to I\!R, |f^{(p)}(x) - f^{(p)}(y)| \leqslant M|x - y|^{\beta - p} \right\}, \tag{18}$$

where p is the largest integer strictly less than β. Let $\beta_1 > \beta_2 > \cdots > \beta_k > 0$. Set $\mathcal{F}_i = F(\beta_i, M)$ and as in Section 3 let $\mathcal{G}_j = \cup_{i=1}^{j} \mathcal{F}_i$.

Now suppose that we observe the white noise with drift process (5) and that the linear functional is point evaluation where for convenience we take $Tf = f(0)$. Straightforward calculations show that for $i \geqslant j$ there are constants $C_{i,j}$ and C_j such that

$$\omega(\varepsilon, \mathcal{F}_i, \mathcal{F}_j) = C_{i,j}\varepsilon^{\frac{2\beta_i}{2\beta_i+1}}(1 + o(1))$$

and

$$\omega(\varepsilon, \mathcal{G}_j) = C_j\varepsilon^{\frac{2\beta_j}{2\beta_j+1}}(1 + o(1)).$$

It follows that $r(\mathcal{G}_j, \alpha)$ is of order $n^{-\frac{\beta_j}{2\beta_j+1}}$.

Let $\hat{T}$ be the estimator defined in (16). It follows from Theorem 2 that there are constants $C_j(\alpha)$ such that

$$\sup_{f\in\mathcal{G}_j} P(|\hat{T} - Tf| \geqslant C_j(\alpha)n^{-\frac{\beta_j}{2\beta_j+1}}) \leqslant \alpha$$

for all $1 \leqslant j \leqslant k$ and all n. Hence

$$r(\hat{T}, \mathcal{G}_j, \alpha) \leqslant C_j(\alpha)n^{-\frac{\beta_j}{2\beta_j+1}}.$$

In this sense $\hat{T}$ is a rate optimal adaptive estimator over the finite collection of Lipschitz classes under probabilistic error.

This example also illustrates a common situation where estimators which are adaptive under probabilistic error cannot be used to construct corresponding rate optimal adaptive confidence intervals. In fact in this case such adaptive confidence intervals do not exist. For any confidence interval with a given coverage probability of at least $1 - \alpha$ over $\mathcal{G}_k$ the maximum expected length over $\mathcal{G}_j$ must be of order $n^{-\frac{\beta_k}{2\beta_k+1}}$ whereas there are confidence intervals with coverage probability of at least $1 - \alpha$ over $\mathcal{G}_j$ which have maximum expected length of order $n^{-\frac{\beta_j}{2\beta_j+1}}$. See [21,4].

4.2. *Nearly black object*

The nearly black object arises naturally in wavelet function estimation and for estimating the whole object has been studied for example in Donoho, Johnstone, Hoch and Stern (1992) and Abramovich, Benjamini, Donoho and Johnstone (2000). It was also considered in [4,6] for minimax estimation under mean squared error and in the construction of confidence intervals for a linear functional. This example is used here to show that the factor $z_{\alpha/2k} \sim \sqrt{\log k}$ (for large k) in the upper bound in Theorem 2 cannot in general be dropped.

Consider the Gaussian sequence model:

$$y_i = f_i + \frac{1}{\sqrt{n}}z_i \quad i = 1, \ldots, n, \tag{19}$$

where $z_i \overset{iid}{\sim} N(0,1)$ and linear functional $Tf = \sum_{i=1}^n f_i$.

Fix $m_n \leqslant n^\gamma$ where $\gamma < \frac{1}{2}$ and let $\mathcal{G}$ be the collection of vectors in $\mathbb{R}^n$ with at most m_n nonzero entries. Let $\mathcal{I}(m_n, n)$ be the class of all subsets of $\{1, \ldots, n\}$ of m_n elements and for $I \in \mathcal{I}(m_n, n)$ let

$$\mathcal{F}_I = \{f \in \mathbb{R}^n : f_j = 0 \quad \forall j \notin I\}.$$

Note that $\mathcal{F}_I$ is a m_n dimensional subspace spanned by the coordinates in I. These are obviously convex and $\mathcal{G} = \cup \mathcal{F}_I$ where the union is taken over I in the set $\mathcal{I}(m_n, n)$. From now on we shall assume that I is in the set $\mathcal{I}(m_n, n)$.

Simple calculations as in [6] show that for all $I,\ J \in \mathcal{I}(m_n, n)$

$$\omega(\varepsilon, \mathcal{F}_I, \mathcal{F}_J) = \sqrt{\mathrm{Card}(I \cup J)}\varepsilon$$

and consequently

$$\omega(\varepsilon, \mathcal{F}_I, \mathcal{G}) = \omega(\varepsilon, \mathcal{G}, \mathcal{F}_I) = \omega(\varepsilon, \mathcal{G}) = \sqrt{2m_n}\,\varepsilon.$$

Let $K = \binom{n}{m_n}$ be the number of the m_n-dimensional parameter spaces $\mathcal{F}_I$. The following theorem gives a lower bound on the probabilistic error over $\mathcal{G} = \cup \mathcal{F}_I$.

Theorem 3. *Suppose that $n \geqslant 4$ and $m_n < n^\gamma$ with $\gamma < \frac{1}{2}$. Let $Tf = \sum_{i=1}^n f_i$ and let $0 < \alpha < \frac{1}{2}$ be fixed. Then there exists a constant $C(\alpha) > 0$ such that for any estimator $\hat{T}$*

$$\sup_{f \in \mathcal{G}} P(|\hat{T} - Tf| > C(\alpha)\omega\left(\frac{\sqrt{\log K}}{\sqrt{n}}, \mathcal{G}\right)) \geqslant \alpha \tag{20}$$

and consequently

$$r(\mathcal{G}, \alpha) \geqslant C(\alpha)\omega\left(\frac{\sqrt{\log K}}{\sqrt{n}}, \mathcal{G}\right).$$

This result shows that the factor $z_{\alpha/2K} \sim \sqrt{\log K}$ in the upper bound given in Theorem 2 cannot in general be dropped.

5. Proofs

Proof of Theorem 1. We shall focus on the proof for the white noise with drift model (5). The proof for the sequence model (6) is analogous. Fix $\varepsilon > 0$. For any $\delta > 0$ there are functions $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$ such that

$$|Tf_2 - Tf_1| \geqslant \omega_+\left(\frac{2z_\alpha}{\sqrt{n}}, \mathcal{G}\right) - \delta$$

and such that

$$\|f_2 - f_1\|_2 \leqslant \frac{2z_\alpha}{\sqrt{n}}.$$

Denote by P_i the probability measure associated with the white noise process

$$dY(t) = f_i(t)\,dt + \frac{1}{\sqrt{n}}dW(t), \quad -\frac{1}{2} \leqslant t \leqslant \frac{1}{2},\ i = 1, 2.$$

Let $\beta_n = n\|f_1 - f_2\|_2^2$. Then a sufficient statistic for the family of measures $\{P_i : i = 1, 2\}$ is given by the log-likelihood ratio $S_n = \log(dP_2/dP_1)$ with

$$S_n \sim \begin{cases} N(-\frac{\beta_n}{2}, \beta_n) & \text{under } P_1, \\ N(\frac{\beta_n}{2}, \beta_n) & \text{under } P_2. \end{cases}$$

An equivalent sufficient statistic is thus given by

$$Q_n = \frac{Tf_1 + Tf_2}{2} + \frac{Tf_2 - Tf_1}{\beta_n} \cdot S_n,$$

where

$$Q_n \sim \begin{cases} N\left(Tf_1,\ \frac{(Tf_2 - Tf_1)^2}{\beta_n}\right) & \text{under } P_1, \\ N\left(Tf_2,\ \frac{(Tf_2 - Tf_1)^2}{\beta_n}\right) & \text{under } P_2. \end{cases}$$

In testing between $H_0 : \theta = Tf_1$ and $H_a : \theta = Tf_2$ based on Q_n the Neymann–Pearson Lemma yields for any test Γ if

$$P_{Tf_1}(\Gamma = 1) \leqslant \alpha$$

then

$$P_{Tf_2}(\Gamma = 0) \geqslant \alpha.$$

Any estimator $\hat{T}$ yields a test by setting $\Gamma = 1$ when $\hat{T} \geqslant \frac{Tf_1 + Tf_2}{2}$ and $\Gamma = 0$ otherwise. Therefore

$$\max_{i=1,2} P_{Tf_i}\left(|\hat{T} - Tf_i| \geqslant \frac{1}{2}\omega\left(\frac{2z_\alpha}{\sqrt{n}}, \mathcal{G}\right) - \frac{\delta}{2}\right) \geqslant \alpha.$$

The Theorem follows on taking the limit as $\delta \to 0$. □

Proof of Theorem 2. We shall assume that $f \in \mathcal{F}_j$ and show that

$$\sup_{f \in \mathcal{F}_j} P_f\left(|\hat{T} - Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)\right) \leqslant \alpha. \tag{21}$$

The theorem is then an immediate consequence of (21).

Note that

$$\begin{aligned} &P_f\left(|\hat{T} - Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)\right) \\ &\quad \leqslant \sum_{l=1}^{j} P_f\left(|\hat{T}_l - Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right) \cap 1(\Lambda_l = 1)\right) + P(\Lambda_j = 0) \end{aligned}$$

$$\leqslant \sum_{l=1}^{j} \min\left\{ P_f\left(|\hat{T}_l - Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)\right), P(\Lambda_l = 1)\right\}$$
$$+P(\Lambda_j = 0). \tag{22}$$

Let $1 \leqslant l \leqslant j$. There are two cases. In the first case consider

$$|E\hat{T}_l - Tf| > \frac{11}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right).$$

Then either $E(\hat{T}_l - Tf) > \frac{11}{2}\omega(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j)$ or $E(\hat{T}_l - Tf) < -\frac{11}{2}\omega(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j)$. We only consider the first of these since the argument for the second is analogous.

Note that

$$E(\hat{T}_{j,l} - Tf) \leqslant \frac{1}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)$$

and

$$Var(\hat{T}_l - \hat{T}_{j,l}) \leqslant \frac{4}{z^2_{\alpha/2k}}\omega^2\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right).$$

Hence

$$\begin{aligned} P(\Lambda_l = 1) &\leqslant P\left(\hat{T}_l \leqslant \hat{T}_{j,l} + \frac{3}{2}(\omega_{j,l} + \omega_l)\right) \\ &\leqslant P\left(\hat{T}_l \leqslant \hat{T}_{j,l} + 3\omega(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j)\right) \\ &\leqslant P\left(Z \geqslant \frac{2\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)}{\frac{2}{z_{\alpha/2k}}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)}\right) \\ &\leqslant \frac{\alpha}{2k}. \end{aligned} \tag{23}$$

The second case is when

$$|E\hat{T}_l - Tf| \leqslant \frac{11}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right).$$

Note that since $Var(\hat{T}_l) \leqslant \frac{1}{z^2_{\alpha/2k}}\omega^2\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)$ standard calculations yield

$$P_f\left(|\hat{T}_l - Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)\right) \leqslant \frac{\alpha}{2k}. \tag{24}$$

It then follows from (23) and (24) that for $f \in \mathcal{F}_j$

$$\min\left\{ P_f\left(|\hat{T}_l - Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)\right), P(\Lambda_l = 1)\right\} \leqslant \frac{\alpha}{2k}. \tag{25}$$

Now consider $P(\Lambda_j = 0)$. Note that

$$P(\Lambda_j = 0) \leqslant \sum_{l=j+1}^{k} \left(P\left(\hat{T}_j < \hat{T}_{j,l} - \frac{3}{2}(\omega_{j,l} + \omega_j)\right) + P\left(\hat{T}_j > \hat{T}_{l,j} + \frac{3}{2}(\omega_{l,j} + \omega_j)\right)\right).$$

Note also that

$$E\left(\hat{T}_j - \hat{T}_{l,j} - \frac{3}{2}(\omega_{l,j} + \omega_j)\right) \leqslant -(\omega_{l,j} + \omega_j)$$

and

$$Var(\hat{T}_j - \hat{T}_{l,j}) \leqslant \frac{1}{z^2_{\alpha/2k}}(\omega_{l,j} + \omega_j)^2.$$

Hence

$$P\left(\hat{T}_j > \hat{T}_{l,j} + \frac{3}{2}(\omega_{l,j} + \omega_j)\right) \leqslant \frac{\alpha}{2k}.$$

Similarly

$$P\left(\hat{T}_j < \hat{T}_{j,l} - \frac{3}{2}(\omega_{j,l} + \omega_j)\right) \leqslant \frac{\alpha}{2k}.$$

Hence

$$P(\Lambda_j = 0) \leqslant \frac{(k-j)\alpha}{k}. \tag{26}$$

It follows from (25) and (26) that

$$\sum_{l=1}^{j} \min\left\{ P_f\left(|\hat{T}_l - Tf| \geqslant \frac{13}{2}\omega\left(\frac{z_{\alpha/2k}}{\sqrt{n}}, \mathcal{G}_j\right)\right), P(\Lambda_l = 1)\right\} + P(\Lambda_j = 0) \leqslant \frac{j\alpha}{2k} + \frac{(k-j)\alpha}{k} \leqslant \alpha$$

and the theorem follows. □

Proof of Theorem 3. In the proof write m for m_n. Let ψ_μ be the density of a normal distribution with mean μ and variance $\frac{1}{n}$. And for $I \in \mathcal{I}(m, n)$ let

$$g_I(y_1, \ldots, y_n) = \prod_{j=1}^{n} \psi_{f_j}(y_j),$$

where $f_j = \frac{\rho}{\sqrt{n}} 1(j \in I)$ and $\rho = \sqrt{\frac{1}{2}\log\frac{n}{m^2}}$. Finally let

$$g = \frac{1}{\binom{n}{m}} \sum_{I \in \mathcal{I}(m,n)} g_I$$

and $f = \prod_{j=1}^{n} f_0$ be the density of n independent normal random variables each with mean 0 and variance $\frac{1}{n}$. A similar mixture prior was used in [1] to give lower bounds in a nonparametric testing problem.

Note that for all g_I, $Tg_I = m\frac{\rho}{\sqrt{n}}$ and that for any $d > 0$ if

$$P_{g_I}\left(|\hat{T} - Tg_I| > d\right) < \alpha$$

for all $I \in \mathcal{I}(m, n)$ then it follows that

$$P_g\left(\left|\hat{T} - m\frac{\rho}{\sqrt{n}}\right| > d\right) < \alpha.$$

In [4] it is shown that when $n \geqslant 4$ and $m < n^{\gamma}$ with $\gamma < \frac{1}{2}$ then the L_1 distance between f and g can be bounded by

$$\int |g - f| \leqslant \left(4^{n^{-(1-2\gamma)}}\left(1 + \frac{1}{n^{\frac{1}{2}}}\right)^{n^{\gamma}} - 1\right)^{\frac{1}{2}} \downarrow 0.$$

Hence for any $0 < \varepsilon < 1 - 2\alpha$ there exists n_ε such that for all $n \geqslant n_\varepsilon$, $\int |g - f| \leqslant \varepsilon$.

Let $\hat{T}$ be an estimator of T. Any such estimator can be used to construct a test between $H_0 : \theta = Tf$ and $H_a : \theta = Tg$ as follows. Choose H_0 when $\hat{T} - Tf \leqslant \frac{m\rho}{2\sqrt{n}}$ and choose H_a otherwise. Note that

$$P(\text{Type I error} + \text{Type II error}) \geqslant 1 - \frac{1}{2}\int |g - f| \geqslant 1 - \frac{\varepsilon}{2}.$$

See for example [18]. Hence if

$$P\left(\hat{T} - Tf > \frac{m\rho}{2\sqrt{n}}\right) < \frac{1}{2}$$

then

$$P_g\left(\hat{T} - Tg < -\frac{m\rho}{2\sqrt{n}}\right) > \frac{1}{2} - \frac{\varepsilon}{2}.$$

Hence

$$\max\left\{P\left(|\hat{T} - Tf| > \frac{m\rho}{2\sqrt{n}}\right), P_g\left(|\hat{T} - Tg| > \frac{m\rho}{2\sqrt{n}}\right)\right\} \geqslant \frac{1}{2} - \frac{\varepsilon}{2}.$$

Note that the number of convex subsets K is equal to n choose m_n and it is easy to see that

$$K = \binom{n}{m_n} \leqslant n^{m_n}.$$

Hence $\frac{m\rho}{2\sqrt{n}} \geqslant C\omega(\frac{\sqrt{\ln K}}{\sqrt{n}}, \mathcal{G})$ for some constant $C > 0$. Hence the theorem follows. □

Acknowledgments

We thank the referees for very helpful comments which have helped to improve the presentation of the paper.

References

[1] Y. Baraud, Non-asymptotic minimax rates of testing in signal detection, Bernoulli 8 (2002) 577–606.
[2] L.D. Brown, M.G. Low, A constrained risk inequality with applications to nonparametric functional estimation, Ann. Statist. 24 (1996) 2524–2535.
[3] T. Cai, M. Low, On adaptive estimation of linear functionals, Ann. Statist. 33 (2005), to appear.
[4] T. Cai, M. Low, An adaptation theory for nonparametric confidence intervals, Ann. Statist. 32 (2004), 1805–1840.
[5] T. Cai, M. Low. Adaptive estimation of linear functionals under different performance measures, Technical Report, Department of Statistics, University of Pennsylvania, Bernoulli 11 (2005), to appear.
[6] T. Cai, M. Low, Minimax estimation of linear functionals over nonconvex parameter spaces, Ann. Statist. 32 (2004a) 552–576
[7] T. Cai, M. Low, Nonparametric function estimation over shrinking neighborhoods: Superefficiency and adaptation, Ann. Statist. 33 (2005), in press.
[8] D.L. Donoho, Statistical estimation and optimal recovery, Ann. Statist. 22 (1994) 238–270.
[9] D.L. Donoho, R.G. Liu, Geometrizing rates of convergence III, Ann. Statist. 19 (1991) 668–701.
[10] S. Efromovich, Nonparametric Curve Estimation: Methods, Theory and Applications, Springer, New York, 1999.
[11] S. Efromovich, On logaritmic penalty in adaptive pointwise estimation and blockwise wavelet shringkage, J. Comput. Graph. Statist. 14 (2005), 20–40.
[12] S. Efromovich, M.G. Low, Adaptive estimates of linear functionals, Probab. Theory Rel. Fields 98 (1994) 261–275.
[13] S.Y. Efromovich, M.S. Pinsker, An adaptive algorithm of nonparametric filtering, Automat. Remote Control 11 (1984) 58–65.
[14] R.H. Farrell, On the best obtainable asymptotic rates of convergence in estimation of a density function at a point, Ann. Math. Statist. 43 (1972) 170–180.
[15] M. Hoffmann, O.V. Lepski, Random rates in anistropic regression (with discussions), Ann. Statist. 30 (2002) 325–396.
[16] Y.-G. Kang, M.G. Low, Estimating monotone functions, Statist. Probab. Lett. 56 (2002) 361–367.
[17] A.P. Korostelev, V. Spokoiny, Exact asymptotics of minimax Bahadur risk in Lipschitz regression, Statistics 28 (1996) 13–24.
[18] L. Le Cam, Asymptotic Methods in Statistical Decision Theory, Springer, New York, 1986.
[19] O.V. Lepski, On a problem of adaptive estimation in Gaussian white noise, Theory Probab. Appl. 35 (1990) 454–466.
[20] O.V. Lepski, V.G. Spokoiny, Optimal pointwise adaptive methods in nonparametric estimation, Ann. Statist. 25 (1997) 2512–2546.
[21] M.G. Low, On nonparametric confidence intervals, Ann. Statist. 25 (1997) 2547–2554.
[22] A.B. Tsybakov, Pointwise and sup-norm sharp adaptative estimation of functions on the Sobolev classes, Ann. Statist. 26 (1998) 2420–2469.
[23] L. Weiss, J. Wolfowitz, Estimation of a density at a point, Z. Wahrscheinlichkeit. Verw. Gebiete 7 (1967) 327–335.